DOCUMENT RESUME

ED 183 191 IR 008 133

AUTHOR Muiznieks, Viktors; Dennis, J. Richard
TITLE A Look at Computer-Assisted Testing Operations. The Illinois Series on Educational Application of Computers, No. 12e.
INSTITUTION Illinois Univ., Urbana. Dept. of Secondary Education.
SPONS AGENCY EXXON Education Foundation, New York, N.Y.
PUB DATE 79
NOTE 48p.; Best copy available; For related documents, see IR 008 123-139.

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Computer Assisted Instruction; *Computer Managed Instruction; Cost Effectiveness; Guidelines; Item Analysis; Item Banks; Item Sampling; *Program Descriptions; *Student Testing; *Test Construction; Test Items

ABSTRACT

In computer assisted test construction (CATC) systems, the computer is used to perform the mechanical aspects of testing while the teacher retains control over question content. Advantages of CATC systems include question banks, decreased importance of test item security, computer analysis and response to student test answers, item analysis assistance, capability of being integrated into individualized instructional programs, performance of clerical tasks, system availability, and question generation following a prespecified pattern. Major features of CATC systems include item banking, generation, attribute banking, and selection, as well as test scoring, printing, and interactive computer testing in linear and branched formats with sequential and multi-staged approaches. Major limitations of CATC systems include possible lack of lower case letters, character type limitations, lack of pictorial information presentation capability, no backspacing capability, difficulty in printing subscripts or superscripts, and inability to control the quality of item bank materials. Five existing CATC projects are described, and relative costs of CATC are discussed. This resource for preservice and inservice teacher training includes an 85-item bibliography. (CMV)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED183191

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Number 12e

Illinois Series on Educational Application of Computers

A LOOK AT COMPUTER - ASSISTED TESTING OPERATIONS

Viktors Muiznieks

and

J.Richard Dennis

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

J. Richard Dennis

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

College of Education
Department of Secondary Education
University of Illinois
Urbana, Illinois

This work was supported by a grant from
The EXXON Education Foundation

Copyright, 1979, Department of Secondary Education
University of Illinois

IR008133

To the reader,

This is one of a series of documents prepared as resources for the preservice and inservice training of teachers. The general theme of all items in this series is "teaching with or about computers". The booklets are the collective work of a variety of staff and students (both present and former) of the College of Education, University of Illinois at Urbana-Champaign. The project generously was supported between 1977 and 1979 by a grant from the EXXON Education Foundation. All responsibility for the content of these documents rests with the authors and does not reflect necessarily the views of the EXXON Foundation. It was my extreme pleasure to direct this effort throughout.

J.Richard Dennis, PhD
Department of Secondary Education
University of Illinois

A LOOK AT COMPUTER-ASSISTED TESTING OPERATIONS

One of the applications of computers to education which has found rapidly increasing favor since the mid-1960's is the facilitation of various facets of testing and of test construction. In computer-assisted test construction (CATC) systems, the computer is used to perform the mechanical aspects of testing, while the teacher retains control over question content. Lippey (1973, p.11) has stated:

> This (mechanized) role of the computer is easy for instructional personnel to accept, since it does not force change in existing practices. Computer assistance in test preparation can support almost any teaching approach. It can be offered as a service which any educator can elect to use or not use, as he sees fit. It can be set up to operate entirely under his control, supporting him rather than imposing structure.

Computer-assisted test construction still is in its infancy. Neither educators nor system designers yet are certain which of the many possible operations involved in testing or test construction should be executed by a computer, which should remain as manual activites, or which should be omitted altogether.

Computer-based tests generally have been characterized as groups of test items which are stored and retrieved, or generated by means of a computer system. However, the actual implementation o those functions has been extremely diverse. Lippey (1973, p.10) notes that:

> Systems having similar test preparation functions often were developed for different reasons. Systems which were motivated by similar considerations often support somewhat different functions. Some systems restrict themselves to support of test preparation; others offer test preparation as one of several related services, such as test scoring. While most systems are designed to support evaluation or diagnosis, some also provide lists of questions for drill or review (Hsu and Carlson, 1973; Toggenburger, 1973) or for self-evaluation (Sivertson, Hansen, and Schoenberger, 1973).

There appear to be two main influences which are responsible for the widespread diversity in CATC systems (Byrne, 1975). First, there are the needs of the persons or organizations which the system is to serve. Second, there are the particular

interests and areas of expertise possessed by those who design and develop the computer testing system.

The fact that the needs of various organizations and teachers differ is obvious. Some organizations require thousands of copies of standardized tests of very high quality (e.g., Educational Testing Service). However, individual teachers require relatively small numbers of tests that meet their personal teaching objectives, that suit the abilities of their students, or that can be administered in a suitable length of time. The length of time acceptable between requesting and receiving a test (or the scored results) varies from location to location, as does the importance attributed to the security of the test items themselves. Hence, it is common to find descriptions of CATC systems making clear the needs of the users being served.

The contention that CATC systems often reflect the interests and areas of expertise of their designers perhaps is not obvious. Lippey (1973, p.12) claims that CATC systems "are usually stimulated by classroom instructors, rather than professional innovators experienced in applying computer technology to education." However, classroom instructors usually do not have the expertise necessary to produce CATC systems. They must rely on CATC system designers for the implementation of their requirements.

Byrne (1975, p.12) claims that CATC system designers can be classified into three broad groups:

1. The computer-oriented

2. The test-oriented,

3. The teacher(or user)-oriented

Computer-oriented designers tend to produce CATC systems which can be characterized by the possession of many technical facilities. The systems are designed and built with great care so as to maximize the efficiency of the computer programs and the overall system operations. An example of such a system is MEDSIRCH, developed by Hazlett (1970,1973), which has special programs for performing operations such as sequentially revising the item pool and counting the number of records and items in the pool.

Test-oriented designers produce systems which ensure that the range and type of specifications permitted are appropriate for producing tests of the type needed. Much attention is given to the means by which questions are assigned indexing codes. A concept of question (or of test) quality is developed and questions which fail to meet the appropriate standards (either in

design or in operation) will not be retained. An example of such a system is the one developed for Educational Testing Service by Epstein, 1973.

Teacher (or user)-oriented designers place great emphasis on "satisfying" a reasonable number of users. Attention is given to easy and rapid communication with the system, to the speed of assembly of a test, and to the means by which multiple copies of a test readily can be obtained. Such a system usually provides rapid test scoring facilities, and performs analyses of test results in an attempt to diagnose students' strengths and weaknesses for the benefit of both teachers and students. An example of such a system is the Question Pool Management System (QPMS) developed by Denney, (1973).

CATC systems are designed to be computer-oriented, test-oriented, or teacher- (user) oriented based on the interest and expertise of their designers. Some systems are very sophisticated in one aspect of their design, but are more primitive in other aspects. Other systems display a more balanced design. However, regardless of the composition of their design, various systems differ in the attention they give to the particular environment in which they are to operate. Different combinations of design techniques are required to meet the different needs of particular users. In any given situation, the appropriate mixture of design techniques should be determined by the needs of the intended users and the specific environment(s) in which the system is envisioned to operate.

RATIONALE FOR USING COMPUTERS IN TESTING

There are several advantages in using a computer to augment the testing process. A few of the more important ones will be discussed here.

Many test construction systems provide question banks (collections of test questions organized to enable easy access to particular types of questions within the bank) to facilitate the construction of a test from a given specification. The specification might call for a certain number of questions from each of a number of content areas or objectives, or it might also stipulate the difficulty level and intellectual skill(s) which each question should test. This kind of service could be provided by a "hand-operated" bank in which the questions are placed on file cards and indexed via some predetermined scheme.

Occasionally, teachers are interested in specifying overall test characteristics as well as individual question characteristics. For example, a teacher may want to prepare an examination (or two parallel forms of an examination) for which the scores will have certain distribution properties such as a given mean and standard deviation for a given student population. (See Schroeder (1975) for an example of such a system). In such cases, a "hand-operated" system becomes extremely difficult to use and computer assistance becomes desirable.

Question banks also can offer convenient access to large pools of questions. It is possible that the questions have been reviewed for quality (maybe even constructed by skilled testing experts) and evaluated in practice. Therefore, some estimates as to question quality and operating characteristics usually will be known. As Lippey (1974, p.17) states: "The overall quality of an item collection, properly modified as a result of experience, will continue to improve, even though the quality initially might leave much to be desired." The information gained through item analysis (hopefully) will identify deficiencies in particular items. Teachers subsequently can improve the items in a bank thereby making it possible, with much less effort, to obtain tests of higher quality.

Another advantage claimed for large item pools is that the security of test items loses much of its significance. Lippey (1973, p. 11) states "When large numbers of questions become available, the problem of test security (referring to the problem of preventing tests from being seen by future examinees) practically disappears." With several items pooled for each objective, teachers may permit students to experience practice tests without running the risk of items being memorized.

One advantage of using a computer in question banking is the power to control reliably the activities involved in managing a large question bank. A computer can be programmed to facilitate the various operations involved in adding, deleting, or modifying particular items, in identifying questions that seem to be functioning poorly, in locating questions with relatively equivalent characteristics for inclusion on alternate forms of similar examinations, and in controlling access to the question bank itself.

A computer can print a test perhaps with several scrambled versions of the same test, or alternate forms of the tests, on reproduction masters so that multiple copies easily can be obtained. Answer sheets also can be prepared by a properly designed testing system. The production of a test by a computer rather than a typist should yield more uniform results with fewer typographical errors. Once such a system is operational, the relative efficiencies in the process of producing tests should improve instructional settings by encouraging more frequent testing.

CATC systems nearly always have components which analyze and respond to students' test answers either for the benefit of the student or for the teacher. Systems which provide diagnostic feedback and advice to students often go beyond what a teacher is capable of doing in practice. However, such systems do require a means by which the conceptual structure of the course and its objectives are represented and to which the test questions clearly are related.

A computer can be used to assist the teacher in obtaining, analysing, and comparing item/test analysis information. Ebel (1972, p93) writes: "There is no better way for a teacher or professor to continue to improve his skills in testing, and the quality of the tests he uses, than to analyse systematically the results from tests and to compare the findings of these analyses with ideal standards of test quality." However, teachers usually do not analyze their own tests. In some cases it is because teachers do not know how to analyze their tests (Bloom, Hastings, and Madaus, 1971, p 254). With a computer, teachers can receive a complete item/test analysis for every test scored and students quickly can receive test results. It also would seem that using a computer (particularly in large classes) would reduce the error in the correction of (objective) tests and in the computation of scores.

Computerized test-preparation systems are potential components of more sophisticated computer-managed instruction approaches. These more extensive systems can include instructional decision-making elements, such as the diagnosis of learner difficulties and the prescription of individual

assignments. CATC systems also can be integrated into individualized instruction programs. These programs assist students in proceeding through large units of instructional materials independently, without the need for continuous contact with instructional staff. Several existing CATC systems already are components of more extensive computer-managed instruction systems (Denney, 1973; Dudley, 1973; Hsu and Carlson, 1973; Libaw, 1973; and O'Reilly, Gorth, and Pinsky, 1973).

In most CATC systems, the computer performs the clerical tasks associated with the testing process. The computer offers fast, reliable, and accurate service with extensive record-keeping facilities and considerable analytic power. Relatively large amounts of data about specific questions, tests, or students can be stored, easily retrieved, and analyzed for many purposes. As noted by Alessi, Anderson, and Biddle (1975, p.2), "The storage and analysis of large amounts of data are useful not only to instructors but to instructional developers and evaluators for improvement of materials, analysis of overall class progress, and evaluation of work quality." When a computer is used to perform the clerical tasks involved in testing, the more important decisions involved in testing are left to the instructor. As Ansfield (1973, p.13) states, "Time becomes an added and real instructional resource; faculty may spend more time as managers of the learning process, and less as clerks."

Another advantage to using computers in testing is their availability. As Alessi, Anderson, and Biddle (1975, p. 2) note, "Many computing facilities are capable of providing more hours of continuous service than can be provided by the entire instructional staff of the typical university course." Also, a large number of teachers can be using a computer system at the same time.

Still another advantage in using computers in testing is the computer's capability to generate questions following a prespecified pattern (item form). This means that a great deal of savings in time (fewer work-hours for writing questions or more questions in the same amount of time). A greater number of questions covering the same area (or instructional objective), a convenient way of constructing drill-type worksheets, a significant savings in storage space, and a decrease in the amount of time required to specify which questions to include on tests, all can be realized. A more complete discussion of item generation can be found in the next section.

Study Questions

1) Why would a classroom teacher use computer assisted test services?

2) What are the major orientations from which existing CATC systems have been designed? List the advantages and disadvantages possibly implied by each orientation.

A BRIEF SUMMARY OF TESTING

Before describing in more detail the computerization of testing, we shall review for you some of the principles and details of testing itself. This presentation deliberately is brief. Entire courses are presented on this topic alone (see Gronlund, 1976). For our purposes we briefly will describe the purposes of tests, the varieties of learnings we may wish to test, and some of the measures used to determine the effectiveness of a test or a test question.

Why Test

There are several purposes for giving tests in the school setting. Initially, when a student arrives at the school, a means may be needed to determine which grade or which courses the student should enter. In an individual classroom, a teacher may give a test to determine if prerequisite knowledge or skills for the next lesson are present. This is called readiness or placement testing. Sometimes this type of testing will be very comprehensive and broad in the behavior it attempts to measure. Sometimes, as in the case of testing for reading readiness in primary schools, the behaviors tested for may be very specific.

A second purpose for testing in schools is for diagnosis during the learning process. This is a type of questioning conducted in an attempt to pinpoint specific causes of lack of learning or of faulty learning. This is a difficult type of testing to design because one must be aware both of the specific learning errors that can occur and also of how such errors are exhibited in responses to carefully selected questions.

A third purpose for testing is to determine if learning has taken place. This is referred to as pupil progress testing or as mastery testing. This type of testing can be focused at a broad, comprehensive level of a course or a subject or used to determine if individual objectives have been met or mastered.

What Do We Test?

Generally what we test are behaviors and we believe that these are the result of learning. Individual learnings fall in a variety of categories which influence the nature of questioning done to verify that learning. There are several schemes for categorizing learnings. We summarize those presented by Gronlund (1976).

At the most easily identifiable level of learning are the behaviors we call knowledge. There are many examples of knowledge, including vocabulary (whether in a mother tongue or a foreign language) and identifying terminology. Also included are a variety of sets of specific facts such as the arithmetic combinations, telephone numbers, etc. Methods and procedures for carrying out various actions such as calculating usually are categorized as knowledge. In general, knowledge is the set of those behaviors distinguished by recall or by simple identification.

A second category of learning is that which we call understanding. This word gets used in many vaguely defined contexts in education. Among the categories of learning outcomes, however, understanding refers to those behaviors in which awareness of relationships or explanations are exhibited. Sometimes these learnings, also are referred to as concepts, properties, principles, or methods. However, understanding behaviors show more complexity to the learning than simple recall.

At a still more complex level are the learnings referred to as applications. These are demonstrations of the ability to use concepts, priciples, methods, or procedures in a variety of contexts. In a sense, application learnings are demonstrations where we can recognize the appropriateness of knowledge or understandings to help us achieve a specific goal.

The last category of intellectual learning is the one referred to by Gronlund (1976) as thinking skills. This includes the critical, "common sense" type of behavior and the scientific style of thought based upon a mathematical style logic. At this level are the abilities to separate information into component parts and to recognize structures or generalizations in information. Alternately, this category includes abilities to assemble fragments of information into new structures. In a sense, these are one's abilities to arrange and to rearrange concepts, principles, and relationships to solve problems or to create new ones.

Some other categories of learned behavior for which testing occurs are attitudes (both social and those based upon more scientific reasons), interests (both personal and vocational), a wide variety of appreciations (for literature, art, music, achievement), and the various behaviors of personal adjustment.

Effective Test Construction

There are a few principles that can be followed in designing a test. A more elaborate discussion of these principles is found in Gronlund (1976).

1. Prepare the test to agree with the intended use (be it mastery, diagnosis, etc.)

2. Style of the test items used is determined by the levels of learning being tested.

3. Items on a test should represent the broad variety of behaviors included in the instructional goals.

4. Difficulty of the questions should match both the students and the desired learning goals.

5. Questions should be edited carefully so that irrelevant characteristics do not prevent answering.

6. Questions also should be examined to insure that a correct answer has a high probability of indicating learning.

7. Tests are a learning opportunity in themselves and should be constructed and given to contribute to learning.

These considerations particularly are important to computer assisted testing and to the construction or evaluation of questions for such applications.

There are two major types of tests that are common to school applications--the objective test and the essay test. Objective tests are composed of multiple choice, short answer, or matching type items. If the test writer is skilled, objective tests can contain interpretive items written in multiple choice form. It is the objective test that most frequently is involved in computerization. These tests generally require more care and effort to write, partly because a larger number of questions is needed to cover a body of content. Such questions usually are quite constant in their results over classes of comparable ability.

The other style test is the essay test. This style test is best for testing higher level behaviors such as thinking skills. These tests are easy to prepare because single items cover larger portions of the content. For students, essay items place emphasis on organization, integration, and expression of information. However, for computerizing a testing program, essay items are difficult to control. It virtually is impossible to prepare

answers to an essay item for a computer to implement. Human teachers do this very well and we have no replacement for these talents with the present technology.

MAJOR FEATURES OF COMPUTER-ASSISTED TEST CONSTRUCTION

Simply stated, a CATC system is nothing more than an inventory and retrieval system for questions and their related information. A computer can be used to assist test preparation in a variety of ways. Figure 1 indicates the various operations involved in a typical computer-assisted test construction system. The six major features of a CATC system include: item banking, item generation, item attribute banking, item selection, test scoring, and test printing (Lippey, 1974, p. 5). Not every CATC system will provide all of these services.

Item Banking

Item banking refers to the storage of questions in a form accessible and readable by a computer (called a machine-readable form). In some item banks, the questions are stored on magnetic material (such as disks, drums, or tapes), while in others the questions are stored on punched cards. In the latter case, the instructor would assume more of the responsibility for the clerical aspects of keeping the file up to date (adding and deleting items, removing errors, etc.), while the computer would "bank" only such information as question identity numbers and other important identifying characteristics associated with each of the questions.

A computer system typically has one or more means by which information can be entered into the system, and several means of storing information within the system. In creating item banks, two methods for entering information frequently are employed. These are:

1. using previously prepared machine-readable information which is delivered to the computer system in a batch-processing mode, (e.g., punched cards, punched tape, etc).

2. using information which can be entered directly into the computer system using a typewriter-like terminal in an interactive fashion (this may or may not be preceded by a hand scripting or planning of the item information).

When using the first of these methods, items usually are coded first on specially prepared coding sheets that help to organize the various pieces of item information. (See ISEAC no 4e for a detailed discussion of question construction). The information on these sheets later is keypunched (or coded in a mark-sensed form) on computer cards, and many items (via the computer cards) can be read into the computer in one operation. This method has many advantages in initially forming the item pool, and often requires less teacher-time for item entry. In fact, an efficiently designed CATC system should permit clerical personnel to achieve actual question entry. Another advantage of this method is that items can be prepared separately from the computer system.

The second item entry method provides a convenient way to prepare information. It most frequently is used to enter small numbers of questions or to make revisions in existing questions. This method has some disadvantages in that a computer must be available before any information can be entered. Also, a machine-readable copy of the data is not available separately from the computer system without initiating some additional procedures. (Imagine the frustration which might result if, after three or four hours of typing items and their associated attributes, the computer system "crashes" and the user must start over because no "backup" or "off-line" copy of the information had been made.)

In actual practice, a CATC system which offers a combination of these item entry methods probably is desirable. When initially preparing the major portion of an item pool, an educator would want an efficient, off-line method of recording item information and an efficient method for clerical assistants to enter the information into a computer system. For purposes of editing individual items in a pool, or for adding a small number of items to an existing pool, an interactive, teacher operated facility becomes the more efficient mode of operation.

Item Generation

The basic concept underlying automatic generation of test items (called item generation) is that the item-writer merely provides the computer a pattern (or template) from which many test items of the same kind can be produced. It is possible to generate objective questions in this way only when the topic being tested contains an explicit rule structure by which an answer can be determined. Such structures are found frequently in the teaching of mathematics, science, grammer (either in one's native language or in a foreign language), and less frequently in other aspects of the social sciences and humanities.

The patterns from which the computer generates specific items contain two types of elements:

1. constant elements--remain the same for all questions generated from the pattern

2. variable elements--the computer substitutes selections from a defined set of information (either numbers or words) each time a new item is generated.

The choice of variable elements frequently depends upon the application of a random selection scheme. Often, the flexibility and power of that scheme determines the limitations which are imposed on the kinds of item patterns that can be specified.

The simplest types of problems which can be generated from templates are those which merely employ substituting numbers or single words for specified variables. For example,

From a basket full of X apples, John ate Y apples.
What percent of the apples in the full basket did John eat?

In this instance, the computer simply would calculate and supply numbers in place of the X and Y at the time the item was printed on the test. Of course, reasonable restrictions on the size and type of the numbers put in place of X and Y may be necessitated depending upon the student audience and their stage in this learning situation. The computer system should be able to aid the item-writer in this regard. Saying the value of Y should be less than the value of X must be easily achieved. Being able to specify that the value of X and the value of Y are both integers also may be important in a particular context, as well as being able to define an acceptable range for their values (e.g., X is between 30 and 50, and Y is less than 5 but greater than 0).

At a more sophisticated level, the template described above might be prepared like this:

From a(n) A$ of X B$, C$ D$ Y B$.
What percentage of the B$ in the A$ did C$ E$?

In this template, the variables would be defined as follows:

A$ holds a name of a container or other quantity grouping (e.g., basket, bottle, team, fabric bolt, truck load, etc.)

B$ holds a name of a unit of measure (e.g., apples, ounces, players, yards, tons, etc.).

C$ holds a person's name (for purposes of personalizing item these could be the names of the students in the class for whom the test is being designed).

D$ holds an action verb in the appropriate tense (e.g. ate, poured, picked, bought, unloaded, etc.).

E$ holds the appropriate tense of the same action verb (e.g., eat, pour,pick, buy, unload, etc.).

X holds a numerical value.

Y holds a numerical value.

This example should illustrate the degree of flexiblity one can build into an item generating template with a little thought. Such generality in a template frequently results in items that cease to have such obvious similarity with one another. For either practice situations or for testing situations, this variation in items tends to reduce the boredom or "drudgery" of the activity.

An illustration of a more complicated type of specification relating the variables in an item template is the following. Students are to receive practice in factoring quadratic equations of the form:

$$A(X \times X)+BX+C$$

with the restrictions that A is a prime number, C is an even integer greater than 0, and B=(AxC)/2+2. In this case, all of the problems would be factorable, into the form (AX+2)(X+C/2). Variations on this theme easily are imaginable. The only constraint lies in the number and variation of the specifications permitted by the computer test system

Of course, the value of item generation is not limited to mathematics. Consider the following question, taken from the Computer Assisted Testing System (CAT) at the University of Pittsburg. (Variable parts of the question are denoted by "...".)

If ... ml. of ... molar ... is mixed with ... ml. of ... molar ..., the final solution will be:

a) ... molar in ...
b) ... molar in ...
c) ... molar in ...
d) ... molar in ...
e) ... molar in ...

In this example several features are illustrated. The first is that the answers, as well as the questions, can be changed. Therefore, it might be possible to detect when students make common errors without doing a lot of work every time the question is used. In this case, what will happen is that a random number will be generated, a value will be determined based on the random number, an item consisting of the amounts and concentrations of two compounds will be chosen from a prespecified list based on the generated random number, and the answer and distractors will be calculated (also based on the initial random number). The generated answers will be randomly arranged (hence, it is possible to construct similar forms of the test merely by resequencing the alternatives),

and a typical item might come out looking like:

If 43.6 ml. of 1.50 molar NaOH is mixed with 38.5 ml. of 1.14 molar HCl, the final solution will be:

a) 1.33 molar in OH-
b) 1.10 molar in H+
c) 0.26 molar in H+
d) 1.33 molar in H+
e) 0.26 molar in OH- (correct answer is e)

Of course, generating these kinds of items involves a more complex testing system. It takes more time and effort to specify good distractor templates based upon common error types. For such reasons, item-generating schemes often will produce only items that are of the "supply the right answer" rather than "select the correct answer" type. In such cases, it still is possible to evaluate the responses, looking for simple spelling errors or allowing numeric answers to fall within a specified range.

It also should be noticed that it is not just numbers alone that are being generated, but that phrases are chosen from predefined lists as well. This is an approach which can be used in many subjects other than chemistry. If the lists consist of commonly used terms and expressions, then many questions can make use of the same list (and thereby save considerable storage space).

The field of item generation offers a new and exciting approach to test construction. Its benefits are not limited to numeric-type exercises but extend far beyond that to all disciplines in many ways. There is no existing theory as to the "best" way in which item generation should be approached. Indeed, there is a need for increased ingenuity and creativity on the part of item-writers and system designers alike. With item generation, the creativity of the item-writer can be amplified by the power, speed, and accuracy of the computer to produce unlimited numbers of unique, yet relevant, exercises (and their answers). Item generation is an area in which both man and machine must get involved and work together in order to produce high-quality and useful results. With a little practice, teachers can become quite adept at the design of item templates.

Item Attribute Banking

Most computerized test construction systems permit the storage of item attributes (or characteristics). There are several reasons for retaining attributes of questions in a readily available form:

1. as a basis for choosing questions for a particular test,
2. for assistance in estimating the statistical properties of a test,
3. for assistance in interpreting test results,
4. for guidance in identifying items that might need revision,
5. for convenience in relating questions to pertinent ancillary materials.

Item attributes typically are classified into one of two categories:

1. assigned attributes--attributes which are intrinsic to, or implicit in, the question itself;
2. measured attributes--attributes which are dependent upon data collected from actual use of the item.

Some typical assigned attributes include:

identification number	estimated item difficulty (e.g., easy)
the question itself	location in the curriculum
the right answer(s)	keywords identifying item content
the wrong answer(s)	guidelines for grading the question
the subject matter	comments about the question
the taxonomic level	references to ancillary materials
name of item author	instructional objectives being measured
question type	cross-referencing characteristics

Among the most frequently encountered assigned attributes are those established according to subject matter classification. These might be termed topic headings, behaviorally-stated objectives, keywords, etc. The reason they are stored is, in general, to aid in choosing questions for particular tests. Other attributes often are stored for this purpose as well, including information such as item type, taxonomic level, and related material necessary for answering the item. Often attributes such as guidelines for grading the answer (in the case of constructed response items), the correct

answer, and the distractors are kept separate from the text of the item itself. Some items may have to be used with, or to the exclusion of, other items (because, for example, the text of one item gives the answer to another item). Hence, some kind of cross-referencing attributes might be employed.

Some typical measured attributes include:

difficulty index (possibly for different uses of the item)
discrimination index (again, might depend on type of use)
effectiveness of distractors (depending on item use)
number of times item used (on system or by particular people)
number of times item looked at but rejected (for system or person)
last time item used (for system or person)
previous version of item (modification history)
time required to answer item (for different purposes)

Measured attributes typically include an index of discrimination and a difficulty index for each item. Separate indices may be kept for different uses of the same item i.e., in formative, placement, diagnostic, or summative testing (or the indices may be computed separately each time the item is used and not accumulated). For multiple-choice items, a record of the effectiveness of each alternative usually is kept. Also, data on item popularity and usage may be valuable for periodic review or culling of the item pool. Attributes of this kind would include a record of the date on which the item was last used, to whom it was administered, and the frequency with which the item should be released from the item bank.

The number of attributes used varies from system to system. The specific attributes found in a particular system reflect the objectives of the system designers. Hopefully, consideration is given to the relationship between the information stored and its eventual utility to a system user, or to a person in charge of maintaining the item bank. Item data should be kept only for those characteristics which need to be analyzed. However, it often is a difficult task to determine exactly what the needs of a system will be at some future date. Hence, the expandability of a system is important. Often users are not informed of the full range of capabilities present in a system until they are at a point where they need to know that information. Attributes need not always be constant. Some flexibility often is desirable in permitting teachers to modify the general classification (name) of an attribute. Frequently, the capability of adding new attributes to an item pool is needed.

Item Selection

When the computer assists in item selection, the teacher usually specifies the desired attributes for the test, and the computer selects the questions accordingly. Selection should be possible on the basis of any of the item attributes (or any combination of item attributes) stored within the pool. Usually, however, teachers choose to select items on the basis of only one important attribute (Stodola, 1974, p. 85). Selection can be done in either an on-line, interactive fashion, or an off-line, batch processing fashion.

Typically there is a difference between the number of items requested and the number available that meet the specified criteria. If more items are available than requested, the computer randomly (or on some other basis) can select which items to include in the test. If there are too few items, the user may need to relax some of the specified item selection criteria. In those CATC systems where the computer generates questions from templates, the test specifications determine which templates are to be used.

Some rather sophisticated variations of item selection have been developed, including techniques which optimize some attribute of a test for a particular situation or audience, such as validity, reliability, or internal consistency (see, for example, Epstein, 1973, and Schroeder, 1975). Caution should be exercised when using these optimizing techniques and care always should be taken to ensure that all of the assumptions made by the optimizing programs are met by the intended audience. In the interactive mode, it also is possible to examine, modify, or alter the items selected for any test. The test author could then examine the test item attributes resulting from such a change. A useful feature of a computerized testing system is the ability to print many different "scrambled" forms of the same, or alternate, test(s). Thus, the teacher is able to specify the number and kind(s) of scrambled forms of the test that are desired and they automatically are prepared by the computer.

In selecting items for inclusion in a norm-referenced test, the teacher probably is most interested in test reliability, item difficulty, discrimination level, and correlation with other accepted criteria as well as the subject matter content of the item and the instructional objective being measured. In criterion-referenced testing, the instructional objectives (and the subject matter) are probably the most important attributes with which to be concerned.

No matter how sophisticated an item selection system is or how many options are available to the test constructor, it should be designed so that it provides a very simple test specification procedure perhaps as a subset of a more complex procedure. Extreme

care should be taken to make the procedure as "error-free" as possible so that teachers almost are "guaranteed" trouble-free use of the system.

Test Scoring

When scoring a test, the teacher ideally is seeking information about each of the questions used, about the students, and about the test. Most CATC systems offer the teacher the option of having tests scored by the computer. In order to use this option, the students must either mark their answers on some medium that the computer later can read (such as mark-sensed cards or optically scannable sheets), or take the test on-line (in an interactive mode). The items graded by such systems usually are limited to objective question types such as multiple-choice, alternative-response (e.g., true-false), or matching questions. Some systems use "clever" techniques to get the computer to grade fill-in-the-blank questions (e.g., the student is asked to specify only the "third letter" of the correct answer). A few systems (Bishop, 1970) are attempting partially to grade essay-type questions by computer. At this time, the computer efficiently can store, retrieve, generate, and print essay-type items. Guidelines for scoring essay-type items also can be incorporated into the computerized testing system. However, the complete computer-scoring of open-ended question types (essays, proofs) is a long way from fully being realized.

When scoring tests, a computer also should produce item/test analysis data and, hopefully, would provide guidelines in interpreting such data. The effectiveness of each test item can be determined by analysing the pattern of pupils' responses to the item. Items or tests can be analysed qualitatively, in terms of their content and form, or quantitatively in terms of their statistical properties. Qualitative analysis of tests would include consideration of content validity (how well the test measures the subject-matter content and behaviors under consideration), and such evaluation of items would include assessment of effective item writing procedures. Quantitative analysis principally includes the measurement of the item difficulty and item validity (the relationship between response to the item and the criterion for performance).

Gronlund (1976) indicates that item analysis usually is designed to answer questions such as the following:

1. Did the item function as intended? (e.g., did norm-referenced test items adequately discriminate between high and low achievers?)
2. Were the test items of appropriate difficulty?
3. Were the test items free of defects?
4. Was each distractor effective (in multiple-choice items)?

The answer to these questions can lead to a number of benefits for the learning situation, including:

1. improving individual test items (in selecting or reviewing items for future use)
2. providing a basis for efficient class discussion of test results (easy items can be omitted, and difficult items can receive special emphasis)
3. providing a basis for remedial work (particular areas of weakness can be discovered, for each student or the entire class)
4. providing a basis for the general improvement of classroom instruction (through the identification of material that is consistently mismatched with pupil performance)
5. increasing skill in test construction (by obtaining experience in reviewing and revising inadequate test items)

Of course, traditional item/test analysis techniques are difficult (if not impossible) to apply in situations where students do not receive the same test (e.g., when students receive different "scrambled", "tailored", or "generated" test forms). Currently there are no good, generally agreed upon techniques for performing item/test analysis on individual tests.

If the teacher is using a criterion-referenced testing approach, in which student achievement is measured against predefined standards (objectives) instead of the performance of other students on the same test, then a CATC system can inform the teacher (and student) which objectives the student has mastered and on which objectives the student needs further work. A CATC system further can be used to refer students to materials which will aid them in mastering unlearned objectives. However, such a CATC system requires prespecified relationships among questions, objectives, and remedial materials. Such a comprehensive CATC system requires a great deal of time and effort to construct and to maintain.

Test Printing

Regardless of whether items are retrieved from an item bank or whether they are generated by the computer, there are several ways in which tests can be printed. The most common approach is to print tests directly on reproducible masters. A few CATC systems produce tests by reprinting the same exam many times on a high-speed line printer or terminal. Regardless of the manner in which tests are printed, there are restrictions on the special symbols which the tests can contain (discussed more completely in the next section). In essence, the characters which can be used in printed test questions are limited to those found on a standard typewriter keyboard. Item attribute systems offer a solution to this limited character set. In such attribute banking systems, special symbols, graphs or illustrations can be included in the items since all material is housed outside of the computer.

One service which CATC systems offer is the ability to print several copies of a single test with the items rearranged at random (scrambled). The ordering of the alternative answers to questions can be scrambled as well. For each printed version of the test, a CATC system should print out and store an answer list (appropriately labelled).

Interactive Computer Testing

An almost universal characteristic of CATC systems is the facility for off-line (independent of the computer) administration of the tests. This is due to the desire to test larger groups of people at one time, thus decreasing the capital investment in interactive terminals. However, recent attempts have been made at making interactive test administration possible. An interactive testing situation uniquely and dynamically may be structured for the individual student. In interactive testing, the student's response to each question immediately is evaluated in some way, a testing decision may be made, feedback to the student may be provided, and another item may be selected for presentation to the student (Thomas, 1976, p. 334). The interactive process is repeated as often as necessary until the testing session ends with a determination of student performance.

A distinction needs to be made between interactive testing and other forms of instructional use of computers. In most modes of instructional computing, questions play a major role. Drill and practice programs, for example, consist almost entirely of questions; tutorial programs intersperse questions with text; simulation and inquiry systems use questions as parts of an unfolding dialogue. The interactive test primarily is used to provide useful guidance between units of instruction, whereas the other instructional computing styles generally provide guidance within the units (Thomas, 1976, p.334).

The interactive testing systems which have been developed may be classified into four general categories (Thomas, 1976, p. 334): linear, branched or tailored, sequential, and multi-staged.

The linear interactive test simply is a computerized version of the conventional pencil-and-paper test format. Items are presented to the student in a predetermined order. The correctness of a response to an item is determined through a comparison to a key, or to a generated answer if the item was produced by an item template. Linear interactive tests are common in computer-managed instruction contexts. Each student experiences exactly the same environment except, possibly, for responses to personal errors.

Tests prepared in the branched format have the form of a network of questions. The next item selected for presentation depends on the answer to the previous question(s). Typically, the next item selected is determined on the basis of item difficulty. For norm-referenced testing it is desirable that items function at a 50-percent difficulty level for the group taking the test. For interactive testing it is desirable to produce a test that has about a 50-percent difficulty level for the individual. The items in the pool are calibrated to determine their difficulty levels (for the type of students who will use them) before they are used. A student typically begins by being asked a question of moderate difficulty. A specific test typically will be designed to contain a fixed number of items for each examinee (although the items themselves may differ).

Branched (sometimes called adaptive or "flexilevel") tests successfully have been employed in non-computerized testing settings. Lord (1971) describes a method in which a modification of conventional test administration and item arrangement procedures can force a matching of item difficulty with ability level for each examinee. Although each examinee answers different sets of items, the scoring method provides scores comparable among all examinees. Also, this scheme can serve to minimize testing times by reducing the number of items presented to each examinee.

The general idea of Lord's procedure is as follows. The test items (an odd number of them) are arranged in order of difficulty (from easiest to hardest). The examinee starts with the middle item in the test and proceeds by taking an easier item each time an answer is wrong or a harder item each time an answer is correct. The examinee completes the test after answering one more than half of the total number of items. The scoring of the test and the revealing of correct and incorrect answers is facilitated by a specially prepared answer sheet (see Lord, 1971, for further details). Such a scheme immediately lends itself to a computer testing system.

Sequential testing procedures usually are implemented in the following way (Thomas, 1975, p.35). The student is administered a test item and the response to the item is evaluated. The examinee tentatively is assigned to one of a number of classes representative of the ability estimate of the examinee. The risk of incorrectly assigning the examinee is evaluated relative to the amount of information available at that point in the test, taking into account any parameters specified by the test administrator. If the risk of assignment to an incorrect class is acceptably low, the test will be terminated. If, on the other hand, the risk exceeds an acceptable level, another item is administered, and the decision process is continued.

There are several important differences between the sequential and branched tests. In the sequential test, the number of items administered is variable. Also, the items are presented in a fixed order and each student receives the same (or similarly generated) items. Branched tests are designed to obtain an estimate of the student's ability level on a continuous scale. Sequential tests are designed to classify the student into a group representative of his/her ability level (e.g., high-pass, low-pass, fail).

The multi-staged testing approach requires two or more sub-strategies on sub-tests. The first stage consists of administering a routing test which is used to compute a general level of ability. This level broadly might be characterized as high-middle-low, or specified in greater detail. The second stage consists of a "measurement" test consisting mainly of items at the ability level indicated by the preceeding routing test. An important concern to designers of multi-staged tests has been the routing test. Various approaches to designing the routing test have been explored by Cleary, Linn, and Rock (1968).

MAJOR LIMITATIONS IN COMPUTER-ASSISTED TEST SYSTEMS

Even though computerized test systems set no particular limitation on the content of questions, (and hence are not limited to any particular subject area), they do involve using a computer. A computer can limit the formats which can be used in writing items in several ways, including:

1. some computer systems do not have lower case letters (consider what this might mean for chemical names, NaCl vs. NACL);

2. the characters which a computer can print usually are limited to those typically found on a standard typewriter (hence no square root symbol or Cyrillic alphabet, for example);

3. pictorial information (charts, graphs, pictures, maps, cartoons, etc.) may have to be supplied from a source other than the computer;

4. there may be no provision for being able to backspace in a line (hence, key words or phrases might not be underlined);

5. it is difficult to print subscripts or superscripts.

Another limitation imposed by existing CATC systems is an inability to control the quality of materials maintained by the bank. This problem is due to the lack of an adequate concept of quality in relation to testing materials (Byrne, 1975, p. 16). Without such a concept, it is difficult (if not impossible) to develop high-quality question banking. As Denney (1973, p. 20) points out, "no matter how great the management system may be, if it collects garbage, the image of technology suffers and deters the progress of education." Denney's own experience in operating a CATC system has shown that almost any teacher who has less professional skill than the authors of the materials stored in the computer system will use those materials but seldom will a teacher with a broader background.

This suggests, of course, that an optimal design for a CATC system is one that encourages the individual teacher to be the materials author. One potential advantage to computerized testing systems is the ability systematically to improve a teacher's testing program through the efficient provision of test data, assistance wherever possible in test data interpretation, and

optimizing of regular changes to that teacher's test item pool. In the presence of such services, the teacher's initial skill at question design is much less of an issue. Test pool quality can grow with the teacher, a situation too infrequently realized in the conventional classroom.

STUDY QUESTIONS

1) List the major features of a typical CATC system.

2) Discuss how each major feature of CATC would be useful to a classroom teacher.

3) Write a question generating pattern (using the "template" style) which might be used for constructing a family of questions in your field of study. Give one or two specific items generated by this template. Generalize the template in an attempt to derive one that generates varying items.

4) What are the usual methods of entering test items into a pool? Discuss the advantages or disadvantages of each.

5) Discuss the flexiblity or inflexibility found when using computers to construct tests.

6) Discuss the role of the classroom teacher in CATC systems.

7) Discuss the possible advantages or disadvantages of each of the interactive testing strategy designs.

EXISTING COMPUTER-ASSISTED TEST CONSTRUCTION SYSTEMS

The largest number of CATC systems may be found in post-secondary educational institutions (Lippey, 1974, p. 15), while a smaller number of systems operate in elementary and secondary schools (Denney, 1973; Hsu and Carlson, 1973; O'Reilly, Gorth, and Pinsky, 1973; Toggenburger, 1973). At least one publisher of nationally standardized tests has begun to enlist computer support (Epstein, 1973), and computer-based test construction services are being offered commercially (Libaw, 1973). A few systems deal with postgraduate certification of medical practitioners (Hazlett, 1973), while others support testing in the military service. Along with the CATC activities in the United States, Britain (Buckley-Sharp, 1970; 1973), Canada (Hazlett, 1973), Germany, and Australia all have efforts underway in this style of computer use.

Because of the number of existing CATC systems (Lippey, 1974, estimates over one hundred CATC systems exist), the following discussion will limit itself to describing only a few of the existing projects: Classroom Teacher Support System (CTSS), Comprehensive Achievement Monitoring System (CAM), Educational Testing Service System (ETS), Question Pool Management System (QPMS), and the Mentor and Examination System (MENTREX).

This discussion will serve to emphasize that, rather than there being a somewhat standard model for CATC systems, there is a set of potentially useful testing related functions of computers and that the blending of these features in any one context very much is a matter of the purpose for which the system is designed.

Classroom Teacher Support System (CTSS)

CTSS (Lippey, Toggenburger, and Brown, 1971; Lippey, 1972; Toggenburger, 1973) is a joint development of IBM's Systems Development Division and the Los Angeles Unified School District. The system originally was developed in 1970 to aid in the teaching of United States history to high school students. The initial item bank (eight thousand items) was prepared in one summer by twenty experienced teachers and three clerks. In two years of initial use CTSS had served approximately three hundred teachers. All the questions in the system are multiple-choice items. The questions are classified by course, by category (subject matter), by difficulty level, by behavior level (knowledge or application of knowledge), by keyword, and by an X-dimension which is used to contain references to ancillary materials. A statistics file is kept for each item which includes difficulty level, discrimination

index, distractor effectiveness, and teacher rejection rate. Teachers are free to include their own (unbanked) questions in tests whenever desired.

Tests can be scored by the system at the teacher's option. The system reports to the teacher a variety of results, including frequency distribution and question response frequencies, as well as scores for individual students. The length of time for returning results is from three to five days.

According to Toggenburger (1973, p. 43), "Teachers rate CTSS as extremely helpful. They state that it saves them time and provides quality materials." Typically, teachers use CTSS every two to four weeks. It takes three to five days for a teacher to obtain a test and three to five days to have the test scored and the results returned. Communications between teachers and the data processing center go through the school district mail which services each school once a day. All requests by teachers are submitted on optically scanned sheets. Teachers can request any number of tests (termed "exercises" in the CTSS system), but each test is limited to at most seventy-five questions.

CTSS is an important example of a CATC system because, among other reasons, it is one of the earliest attempts at large-scale computerized test construction operations. CTSS has been made available to eight high school districts as well as eight colleges and universities.

Comprehensive Achievement Monitoring System (CAM)

The CAM system (O'Reilly, Gorth, and Pinsky, 1973) also uses a number of classifiers for each of its questions. However, in addition to those attributes mentioned in the CTSS system, the CAM system associates an objective with each question. The system designers argue that the banking of objectives is the "logical predecessor" to question banking. Consequently, a user may select a test based on a specification of objectives. The objectives then may be inspected, either instead of, or together with, their associated questions.

A primary concern in the development of the CAM system has been the need to provide a system which will allow the construction of different tests used for different purposes. This concern has lead to a great deal of theorizing about testing by the CAM developers. The designers (O'Reilly, Gorth, and Pinsky, 1973, p.33) suggest that tests must be seen in an evaluative, decision-making context, and that most decisions are classifiable into the following five major types:

1. classification (e.g., assign or eliminate students from a treatment due to scarce resources)

2. summative evaluation on programs (e.g., select treatment A over B on effect or efficiency)

3. formative evaluation on programs (e.g., redesign component A of treatment B to meet specifications)

4. instructional management (e.g., place student in component A of program B; repeat student in A, etc.)

5. curriculum validation (e.g., remove objective C from program A and B at level 1; place at level 2)

The designers argue that most CATC projects take decision-type one as the implicit basis for test development and construction and for the use and interpretation of test data. They are critical of this and suggest that before any attempt is made to assemble a test, the present purpose or function of testing should be determined. In essence, the designers are concerned with guiding the user in making proper use of the facilities provided by the testing system. Often a user is not familiar with the variety of test types and the principles which govern their construction and use.

CAM is based on a criterion-referenced testing approach, centering around decision types two, three, and four. The system can design pre-, post-, and retention tests (the various test types are distinguished by the time at which they are administered during the instructional process). "Item analysis" data can be added to questions. It is reported that the item analysis program is based upon "a developing theory of criterion-referenced item validation" (O'Reilly, Gorth, and Pinsky, 1973, p. 33).

CAM is a part of a much larger operation called the Evaluation Service Center (ESC) which has two basic functions -- test construction and the production of computer generated evaluation reports. Teachers are given twenty hours or more of instruction in the use of the CAM-related ESC services included in the instruction are the following: objective banking, test question banking, maintaining a program structure file, maintaining student registry in courses, and the scheduling of tests. Most of these services involve manual, as well as computer, operations. All of the questions and objectives available in the CAM system are stored away from the computer. The questions and-or objectives are referenced based on the identity numbers selected by the computer. The computer then banks and manipulates these various identity numbers while the specific question information associated with each such number is external to the machine.

Educational Testing Service System (ETS)

The Educational Testing Service has been developing the ETS system (Epstein, 1973) for use in assembling nationally standardized tests. The system greatly has been influenced by test construction specialists and has been developed with great care.

The most important distinction of this system is the attention paid to the indexing of questions. As Epstein (1973, p. 23) states, "The basic classification format allows for any number of four-letter major categories, and within each of these categories up to nine digits, each with 10 levels--a staggering number of possible unique classifications." The way in which the categories are used varies from subject to subject. Some are used to specify subject, topic, and sub-topic distictions. Others are used for taxonomic purposes, for identifying references to figures, grade level, etc.

The broad indexing further is supplemented by more specific indexing. Associated with each question is an identity number, up to fifteen twelve-letter keywords, a complete history of the question's uses, up to five complete sets of statistics (including difficulty and discrimination indices), and codes identifying the security level and present activity status of the question.

While all of the question-attribute information is stored in a computer controlled storage devise, the question-text itself is not. ETS requires a large number of test copies and higher quality printing than is economically feasible to be produced by a computer.

The item selection procedure of the ETS system is complicated. However, it is interesting to note that items are selected for inclusion on a test based on the desire to have the remaining pool of stored items as balanced as possible to maximize the number of parallel forms of a test that can be constructed.

Question Pool Management System (QPMS)

QPMS (Denney, 1973) is an example of a limited item-generating system. QPMS has been developed for use in the Shawnee Mission Public Schools (K-12) in Shawnee Mission, Kansas. Many different kinds of questions are managed by QPMS, including the multiple-choice, alternative-response, short answer, matching, and open-ended types. The system allows for the storage of limited graphic material to be used with questions (e.g., dots and

labels can be stored so that they can be joined during the test printing phase to produce figures).

Only multiple-choice questions can be "computer generated." The user must supply a question stem, the correct answer, and up to seven incorrect answers. From these alternatives the computer can generate up to 245 different five-choice questions. If the order of the alternatives is allowed to vary, then 29,400 different questions are possible from the original alternatives.

From a terminal, a teacher can obtain a thirty question test in less than thirty minutes (including the test itself, the student response sheet, the answer key, and diagnostic information for teacher and student).

QPMS is integrated into a larger computer-managed instruction system, the Curriculum Management System (CMS), which produces not only tests but study guides as well. These study guides contain suggested activities and the necessary bibliographies to carry out those activities.

The Mentor and Examination System (MENTREX)

The MENTREX system (Libaw, 1973) is a commercially operated CATC system of considerable complexity. The MENTREX system is available through MENTREX Enterprises, located in Los Angeles, California. Among the features provided by MENTREX are:

1. multiple parallel tests for mastery learning;
2. an analysis of each student's test performance along several dimensions;
3. statements of mastery of learning objectives by an individual;
4. individual tutorial feedback and references to specific information sources which will help a student to learn what it is that the test performance indicates has not been learned.

All MENTREX tests consist of multiple-choice questions. However, questions can be related to drawings, charts, graphs, etc., which are produced away from the computer. The tests can be printed on spirit masters or offset masters. The system prints both upper and lower case characters.

RELATIVE COSTS OF COMPUTER-ASSISTED TEST CONSTRUCTION

It is difficult to determine the cost of any instructional system and CATC systems offer no exception to this situation. No satisfactory general statement can be made because systems differ so widely in their size, purpose, and levels of sophistication. There are very few references to costs in the literature, and those that do exist usually are not very helpful due to the lack of specificity. However, the general picture seems to be a favorable one. Few systems find their financial demands difficult to justify. Failure to report the cost of a system may be due to the fact that the CATC system is just one of several concurrent projects under development, and individual projects may not be costed separately. More usually, however, only certain aspects of a system are costed (e.g., the cost of retrieving and printing a test).

Lippey (1974) states that CATC systems are seldom supported by special funds. Usually the costs are absorbed into the normal budgetary process. Lippey claims this indicates that the value-to-cost ratio is sufficiently high so that CATC can be funded as an integral part of an on-going program. "Compared to many other computer applications in support of instruction, the economic advantage of test preparation is considerable. The cost of operating a system for constructing tests that are to be used later can be very small (Lippey estimates under $1.00 per student per year)." (Lippey, 1974, p. 20).

Hazlett (1970) has developed an approach for determining the cost of the MEDSIRCH system. Each of the main operations involved in operating a bank (e.g., keypunching time per question) is listed, and the time required to perform each operation (in man-minutes) is determined. Similarly, the computer hardware requirements for each program (such as the amount of storage space used for banking an item) are specified, and the number of input/output operations required for running each program, as well as the execution time of each program, are determined. From this information, and a knowledge of salary scales and computer-related costs, it is possible to calculate the cost of the system along a number of different dimensions.

Several system designers describe various portions of their costs. Usually, however, they report only variable costs (e.g., cost for printing a specific test), and neglect the fixed costs (e.g., data entry and data storage). Salisnjak (1973) found that in 1971-1972, 635 pages of test material were prepared using forty minutes of computer time (costing a total of $70). Salisnjak estimates it would have taken about 160 hours of a secretary's time to type this many pages. Neglecting the fixed costs, this

demonstrated a significant savings in using the CATC system. Menne (1973) reports the cost for computer time in retrieving questions and producing a master copy of a seventy-question, eight-page test (on white bond paper) on the Iowa State University (ISU) system to be 75 cents. CTSS and the Automatic Examination Generator system (Ansfield, 1973), both having more extensive facilities than ISU, are two to four times more expensive to use.

Buckley-Sharp and Harris (1970) have estimated the cost of each item (which has been used at least once) in their question bank to be about .015 cents. This figure includes estimates of all costs incurred in basic data preparation (i.e., obtaining the question and placing it in the bank). They further estimate that a fifty-question test would cost around $6.00. Greater usage of questions would lead to some reduction in the cost.

Hence it is clear that the cost of a CATC system is indeed difficult to estimate. Fixed costs tend to be high, and variable costs tend to be low. The more often an item pool is used, the greater the savings become. Therefore, in representing costs, it often is convenient to amortize fixed costs over time.

LIST OF REFERENCES

Ahmann, J. Stanley, Testing Student Achievement and Aptitudes, The Center for Applied Research in Education, Inc., 1962.

Alessi, Stephen M., Richard C. Anderson, Thomas H. Anderson, W. Barry Biddle, Bruce R. Dalgaard, Donald W. Paden, H. Richard Smock, John R. Surber, Edward J. Wietecha, "Development and Implementation of the Computer Assisted Instruction Study Management System (CAISMS)," Preliminary Technical Report, Contract N61339-73-C-0078 with Naval Personnel Research and Development Center and Defense Advanced Research Projects Agency, University of Illinois, Urbana, March, 1973.

Alessi, Stephen M., Thomas H. Anderson and W. Barry Biddle, "Hardware and Software Considerations in Computer Based Course Management, Technical Report Number 4, Laboratory for Cognitive Studies in Education, University of Illinois, Urbana, November, 1975.

Anastasi, Anne, Psychological Testing (fourth edition), Macmillan Publishing Company, Inc., 1976.

Ansfield, Paul J., "A User Oriented Computing Procedure for Compiling and Generating Examinations," Educational Technology, Volume 8, Number 3, 12-13, March, 1973.

Baker, Frank B., "An Interactive Approach to Test Construction," Educational Technology, Volume 8, Number 3, 13-15, March, 1973.

Baker, Frank B., "Origins of the Item Parameters X50 and B as a Modern Item Analysis Technique," Journal of Educational Measurement, Volume 2, 167-168, 1965.

Baker, Frank B., "The Roles of Statistics," In Lippey (ed.) Computer Assisted Test Construction, Educational Technology Publications, 147-169, 1974.

Barukh, Neve, Eli Sher and Rachel Ramraz, "ITANA-III: A FORTRAN IV Program for Multiple Choice Tests and Item Analysis," Educational and Psychological Measurement, Volume 35, Number 3, 683-684, Autumn, 1975.

Bauernfeind, Robert H., Building a School Testing Program, Houghton Mifflin Company, 1963.

Bishop, R. L., "Computer Analysis of Natural Language Text for Style and Content of Instruction in Writing," Proceedings of the 1970 Conference on Computers in the Undergraduate Curricula, The University of Iowa, June, 1970.

Bloom, Benjamin S., J. Thomas Hastings and George F. Madaus, Handbook on Formative and Summative Evaluation of Student Learning, McGraw-Hill, 1971.

Boyle, Thomas A., "Computer-Scored, Programmed Test Monitors Student Progress," Proceedings of ACM SIGSCE-SIGCUE Joint Symposium on Computer Science and Education, 95-99, February, 1976.

Brown, Willard A., "A Computer Examination Compositor for the IBM 360/40," Educational Technology, Volume 8, Number 3, 15-16, March, 1973.

Buckley-Sharp, M. D., "A Multiple Choice Question Banking System," Educational Technology, Volume 8, Number 3, 16-18, March, 1973.

Buckley-Sharp, M. D. and F.T.C. Harris, "The Banking of Multiple Choice Questions," British Journal of Medical Education, Volume 4, Number 1, March, 1970.

Burris, Bradford, "Innovative Test Reporting Systems," Proceedings of the 15th Annual Convention of the Association for Educational Data Systems, 235-239, 1977.

Byrne, Colin J., "Computerized Question Banking Systems," Institute of Educational Technology, The Open University, January, 1975.

Cartwright, Glenn F. and Jeffrey L. Derevensky, "An Attitudinal Study of Computer-Assisted Testing as a Learning Method," Psychology in the Schools, Volume 13, Number 3, 317-321, July, 1976.

Cleary, T. Anne, Robert L. Linn and Donald A. Rock, "An Exploratory Study of Programmed Tests," Educational and Psychological Measurement, Volume 28, Number 2, Summer, 345-360, 1968.

Cohen, Perrin S. and Leila R. Cohen, "Computer Generated Tests for a Student Paced Course," Educational Technology, Volume 8, Number 3, 18-19, March, 1973.

Coulson, Ronald R. and George K. Wood, "Using the High-Speed Computer for Knowledge Test Construction," Journal of Physical Education and Recreation, 28-29, June, 1975.

Denney, Cecil, "There is More to a Test Pool Than Data Collection," Educational Technology, Volume 8, Number 3, 19-20, March, 1973.

Dirkzwager, A., "Computer Based Testing with Automatic Scoring Based on Subjective Probabilities," In Lecarne and Lewis (eds.) Computers in Education, IFIP, North-Holland Publishing Company, 305-311, 1975.

Dudley, Thomas J., "How the Computer Assists in Pacing and Testing Students' Progress," Educational Technology, Volume 8, Number 3, 21-22, March, 1973.

Ebel, Robert L. (ed.), Encyclopedia of Educational Research, The Macmillan Company, 1969.

Ebel, Robert L., Essentials of Educational Measurement, Prentice-Hall, Inc., New Jersey, 1972.

Educational Testing Service, Multiple-Choice Questions: A Close Look, Educational Testing Service, 1963.

Epstein, Marion G., "Computer Assisted Assembly of Tests at Educational Testing Service," Educational Technology, Volume 8, Number 3, 23-24, March, 1973.

Evans/Methuen Educational, Question Banks: Their Use in School Examinations, Schools Council Publications, 1971.

Fremer, John and Ernest J. Anastasio, "Computer-Assisted Item Writing-I (Spelling Items)," Journal of Educational Measurement, Volume 6, Number 1, 69-74, Spring, 1969.

French, Joseph L. and Donald R. Greer, The Effect of Text-Item Arrangement on Physiological and Psychological Behavior in Primary School Children, University of Missouri, 1964.

Gorth, William P., Dwight W. Allen and Aram Grayson, "Computer Programs for Test Objective and Item Banking," Educational and Psychological Measurement, Volume 31, Number 1, 245-250, Spring, 1971.

Gronlund, Norman E., Measurement and Evaluation in Teaching (third edition), Macmillan Publishing Company, 1976.

Hammer, Mark and C. Obert Henderson, "Improving Large Enrollment Undergraduate Instruction with Computer Generated Repeatable Tests," Proceedings of the 1972 Conference on Computers in Undergraduate Curricula, 209-216, June, 1972.

Hastings, J. Thomas, Phillip J. Runkel, Dora E. Damrin, Robert B. Kane and Gerald L. Larson, The Use of Test Results, University of Illinois, 1960.

Hazlett, C. B., "MEDSIRCH: A Computerized System for the Retrieval of Multiple Choice Items," Alberta University, Edmonton, Department of Educational Administration, Ottawa, Canada, Report Number DERS-3-70, August, 1970.

Hazlett, C. B., "MEDSIRCH: Multiple Choice Test Items," Educational Technology, Volume 9, Number 3, 24-26, March, 1973.

Hoffman, Kaaren I. and George D. Lundberg, "A Comparison of Computer-Monitored Group Tests with Paper-and-Pencil Tests," Educational and Psychological Measurement, Volume 36, Number 4, 791-809, Winter, 1976.

Hotvedt, Martyn Olaf, "A Case Study of Standardized Test Use in the Public Schools," Unpublished Ph.D. dissertation, University of Illinois, Urbana, 1974.

Howze, Glenn, "An Interactive Software System for Computer-Assisted Testing," Proceedings of the 15th Annual Convention of the Association for Educational Data Systems, 17-20, 1977.

Hsu, Tse-Chi and Marthena Carlson, "Test Construction Aspects of the Computer Assisted Testing Model," Educational Technology, Volume 8, Number 3, 26-27, March, 1973.

Jackson, Douglas N., "The Relative Validity of Scales Prepared by Naive Item Writers and those Based on Empirical Methods of Personality Scale Construction," Educational and Psychological Measurement, Volume 35, Number 2, 361-370, Summer, 1975.

Jensen, Donald D., "Toward Efficient, Effective and Humane Instruction in Large Classes: Student Scheduled Involvement in Films, Discussions and Computer Generated Repeatable Tests," Educational Technology, Volume 8, Number 3, 28-29, March, 1973.

Keistinen, Josephine, Karl Lindberg and Stephen Hepler, "A Test Generating System Designed to Facilitate Individualized Instruction in Preparatory Mathematics," ACM SIGCUE Bulletin, Volume 11, Number 1, 8-13, January, 1977.

Kominski, Edward S., "A Data Banking System for State Operated Criterion-Referenced Testing Programs," Proceedings of the 15th Annual Convention of the Association for Educational Data Systems, 11-14, 1977.

Lee, Young B., William B. Michael and Robert A. Smith, "A Computer Procedure for the Item Analysis of a Multiple Choice Test," Educational and Psychological Measurement, Volume 32, Number 1, 203, Spring, 1972.

Libaw, Frieda B., "Constructing Tests with the MENTREX Tutorial Testing System," Educational Technology, Volume 8, Number 3, 30-31, March, 1973.

Libaw, Frieda B., "Pedagogical Implications," In Lippey (ed.) Computer Assisted Test Construction, Educational Technology Publications, 159-201, 1974.

Lippey, Gerald, "A Comprehensive Question Retrieval Application to Serve Classroom Teachers," AFIPS Conference Proceedings, FJCC, 633-639, 1972.

Lippey, Gerald, "Bibliography on Computer-Assisted Test Construction," ACM SIGCUE Bulletin, Volume 7, Number 4, 22-26, October, 1973.

Lippey, Gerald (ed.), Computer-Assisted Test Construction, Educational Technology Publications, 1974.

Lippey, Gerald, "The Computer Can Support Test Construction in a Variety of Ways," Educational Technology, Volume 8, Number 3, 10-12, March, 1973.

Lippey, Gerald and Nathan Parthos, "Improvements to a Computer-Assisted Test Construction System," Educational Technology, Volume 16, Number 12, 21-26, December, 1976.

Lippey, Gerald, Frank Toggenburger and D. C. Brown, "Computer Based Exercises Aid Classroom Teachers," AEDS Journal, March, 75-84, 1971.

Lord, F. M., "The self-Scoring Flexilevel Test", Journal of Educational Measurement, Volume 8, 147-151, 1971.

Millman, Jason, "Creating Domain-Referenced Tests by Computer," A paper presented at the annual meetings of the American Educational Research Association and the National Council for Measurement in Education, New York City, April, 1977.

Menne, John W., "Computer Assisted Test Assembly at Iowa State University," Educational Technology, Volume 8, Number 3, 31-32, March, 1973.

Muscat, Eugene J., "Automatic Achievement Test Generation: Requirements of the Secondary Schools," Proceedings of the ACM SIGSCE-SIGCUE Joint Symposium, 113-115, 1976.

Olympia, P. L. Jr., "Computer Generation of Truly Repeatable Examinations," Educational Technology, Volume 15, Number 6, 53-55, June, 1975.

O'Reilly, Robert P., William P. Gorth and Paul Pinsky, "Computer Assisted Test Construction: An Effort Based on an Evaluation Methodology," Educational Technology, Volume 8, Number 3, 32-34, March, 1973.

Palay, Roger M., "The Structure and Use of a Test Generating System Designed to Facilitate Individually Paced Instruction," Proceedings of the ACM SIGSCE-SIGCUE Joint Symposium on Computer Science and Education, 100-103, February, 1976.

Prosser, Franklin, "Item Banking," In Lippey (ed.) Computer Assisted Test Construction, Educational Technology Publications, 29-66, 1974.

Prosser, Franklin, "Repeatable Tests," Educational Technology, Volume 8, Number 3, 34-35, March, 1973.

Prosser, Franklin and D. D. Jensen, "Computer Generated Repeatable Tests," AFIPS Conference Proceedings, Volume 38, 295-301, 1971.

Remondini, David J., "Test Item System: A Method of Computer Assisted Test Assembly," Educational Technology, Volume 8, Number 3, 35-37, March, 1973.

Remondini, David J. and John E. Miller, "A Computerized System for Preparation of Tests in Academic Disciplines," Proceedings of a Conference on Computers in the Undergraduate Curricula, 7.24-7.30, June, 1970.

Richards, James M., Jr., "Can Computers Write College Admissions Tests?," Journal of Applied Psychology, Volume 51, Number 3, 211-215, June, 1967.

Salisnjak, Julian, "Computer Aided Test Preparation: Six Years of Experience," Educational Technology, Volume 8, Number 3, 37-38, March, 1973.

Schnittjer, Carl J. and Carolyn M. Cartledge, "Item-Analysis Programs: A Comparative Investigation of Performance," Educational and Psychological Measurement, Volume 36, Number 1, 183-187, 1976.

Schroeder, Lee L., "A Program System for the Estimation of Characteristics of the Test Score Distribution Resulting from Test Items with Given Statistics," Educational and Psychological Measurement, Volume 35, Number 3, 685-687, Autumn, 1975.

Seely, Oliver Jr. and Van Willis, "Computer Assisted Test Construction by SOCRATES," In Lecarne and Lewis (eds), Computers in Education, IFIP, North-Holland Publishing Company, 721-724, 1975.

Serlin, Ronald C. and Henry F. Kaiser, "A Computer Program for Item Selection Based on Maximum Internal Consistency," Educational and Psychological Measurement, Volume 36, Number 3, 757-759, Autumn, 1976.

Shoemaker, David M. and H. G. Osburn, "Computer-Aided Item Sampling for Achievement Testing: A Description of a Computer Program Implementing the Universe Defined Test Concept," Educational and Psychological Measurement, Volume 29, Number 1, 165-172, Spring, 1969.

Sivertson, Sigurd E., Richard H. Hansen and Adeline O. Schoenberger, "Computerized Test Bank for Clinical Medicine," Educational Technology, Volume 8, Number 3, 38-39, March, 1973.

Stodola, Quentin C., "Item Classification and Selection," In Lippey (ed.) Computer Assisted Test Construction, Educational Technology Publications, 57-118, 1974.

Stodola, Quentin C., "Use of Computer Assembled Tests in the California State University and College System," Educational Technology, Volume 8, Number 3, 40-41, March, 1973.

Thomas, David B., "Varieties of Interactive Computer-Based Testing," Proceedings of the 14th Annual Convention of the Association for Educational Data Systems, 334-337, 1976.

Toggenburger, Frank, "Classroom Teacher Support System," Educational Technology, Volume 8, Number 3, 42-43, March, 1973.

Vickers, F. D., "Creative Test Generators," Educational Technology, Volume 8, Number 3, 43-44, March, 1973.

Wagner, Arthur and Ronald Bleed, "Using a Computer to Support the Testing Program in Audio-Tutorial Biology," Proceedings of the 1972 Conference on Computers in Undergraduate Curricula, 195-198, June, 1972.

Wahlstrom, Merlin Walter, "A Factor Analytic Item-Selection Procedure," Unpublished Ph.D. dissertation, University of Alberta, Edmonton, 1967.

Wentling, Tim L., "A Comparison of Criterion-Referenced and Norm-Referenced Evaluation with Varying Levels of Specificity of Feedback," Unpublished Ph.D. dissertation, University of Illinois, Urbana, 1971.

Wolfe, John H., "Automatic Question Generation from Text- An Aid to Independent Study," Proceedings of ACM SIGSCE-SIGCUE Joint Symposium on Computer Science and Education, 104-112, February, 1976.

Wood, Robert, "Response-Contingent Tests," Review of Educational Research, Volume 43, Number 4, 529-544, Fall, 1973.